CSC 449, HW#4 (Kefu Zhu)

Problem 1

1. LSTM-based Language Decoder

	Score
BLEU_1	0.635
BLEU_2	0.455
BLEU_3	0.297
BLEU_4	0.192
METEOR	0.197
ROUGE_L	0.460
CIDEr	0.691

2. Implement Your LSTM

	Score
BLEU_1	0.621
BLEU_2	0.453
BLEU_3	0.302
BLEU_4	0.195
METEOR	0.194
ROUGE_L	0.465
CIDEr	0.620
SPICE	0.125

3. Compete against Microsoft's AI

(1) Sport

- **Microsoft's caption bot**: a group of young men playing a game of football

- **My caption bot**: a group of people standing on a field with a frisbee

Microsoft's caption bot is better because it successfully recognizes the football while my caption bot sees it as frisbee

(2) Animal

- **Microsoft's caption bot**: a zebra standing on top of a dry grass field
- **My caption bot**: a zebra standing in a field with a large tree

My caption bot is better because althought it fails to recognizes the lion, it at least knows that there are two objects in the image. but the Microsoft's caption bot only identifies one object, zebra

(3) Scenery

- **Microsoft's caption bot**: pink flower is standing on a lush green field
- **My caption bot**: a large building with a clock on it

Microsoft's caption bot is clearly better since it is at least describing a scenery with reseaonble elements (pink, green field). My caption bot totatlly fails to understand this image.

4. New ideas

Rather than using the CNN extracted features as input of LSTM model, we may want to extract features within segmentations (of the image) using CNN separately and use those features as the input of LSTM model.

By doing so, we can not only have higher probability of recognizing the main objects (segmentations) in the image, but also be able to describe details attributes of each object (CNN extracted feature for each segmentation)

Problem 2

1. BLEU (BiLingual Evaluation Understudy)

Definition:

BLEU aims at measuring the similarity between a machine translation and a professional human translation. **A machine translation that is closer to a human translation will yield higher BLEU score.**

The range of BLEU metric is between 0 and 1

In order to achieve such measurement, BLEU is mainly consisted of two compoenents

- A numerical "translation closeness" metric
- A corpus of good quality human reference translations

BLEU uses **modified n-gram precisions** as the metric for evaluation of "translation closeness", which is able to cover two aspects of tranlation: *adequacy* and *fluency*

- Using the same word (unigram) signifies *adequacy*
- Long n-gram matches account for *fluency*

To combine the modified n-gram precisions, BLEU uses the geometric mean of the modified n-gram precisions, which takes account of the exponential decay in the modified n-gram precision as n gets larger

Furthermore, BLEU also penalties machine translation that has improper sentence length (specifically, sentence that is too short) compared to the reference translation. BLEU introduces a multiplicaive **brevity penalty** factor that ideally will be 1.0 when the machine translation has the same sentence length as any reference translation's length.

Limitations

- **Trouble with recall**

  Since BLEU considers multiple reference translations for a single source, each of which may use different word choices, a good candidate translation will only be similar to one of these styles, which lead to low recall rate.

- **Higher score does not guarantee better translation**

- **Geometric mean of n-gram precision does not work well on segement level**

  If the test corpus is too short, only contains few sentences, some n-gram score might be zero which will yield zero BLEU score when computing the germoetric mean.

2. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

Definition

METEOR is based on an explicit word-to-word matching (unigram matching), where words are matched based on their various form (surface form, stemmed form, synonyms, and etc.). METEOR computes a score for such matching by combining unigram-precision, unigram-recall and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference. The score assigned to each individual sentence of machine translation is derived from the best scoring match among all matches over all reference translations.

In order to perform the explicit word-to-word matching, METEOR first creates an **alignment** between machine translation and reference translation, where every unigram in each sentence maps to zero or one unigram in the other sentence. The alignment is incrementally produced through a series of stages, each stage consisting of two different phases

1. An external module lists all possible unigram mappings from between the two sentence
    - The vanilla METEOR supports three modules
        - **"exact" module**: maps two unigrams if they are exactly the same
        - **"porter stem" module**: maps two unigrams if they are the same after Porter stemming
        - **"WN synonymy" module**: maps two unigrams if they are synonyms of each other
2. The largest subset of these unigram mappings is selected such that the resulting set constitues an valid alignment
    - If more than one subset constitutes an alignment, METEOR will picks the set that has the least number of **unigram mapping crosses**

Each stage only maps unigrams that have not been mapped to any unigram in any of the preceding stages. Therefore, the order in which the stages are performed imposes different priorities on the mapping modules employed by different stages.

After the alignment, METEOR computes a score to measure the similarity between the machine translation and the reference translation using the formula below

$$Score = Fmean * (1 - Penalty)$$

(1) Fmean

Fmean is a harmonic mean of unigram precision (P) and unigram recall (R) that is computed by

$$Fmean = \frac{10PR}{R+9P}$$

- Unigram precision (P): the ratio of the number of unigrams in the machine translation that are mapped, to the total number of unigrams in the machine translation
- Unigram recall (R): the ratio of the number of unigrams in the machine translation that are mapped, to the total number of unigrams in the reference translation

(2) Penalty

To take into account *fluency* (longer matches), METEOR computes penalty for each alignment.

All unigrams in the machine translation that are mapped are grouped into the fewest possible number of chunks such that

- Unigrams in each chunk are in adjacent positions in the machine translation
- Corresponding mapped unigrams are also in adjacent positions in the reference translation

The penalty is then calculated by the following formula

$$Penalty = 0.5 * (\frac{\#\ chunks}{\#\ unigrams\_matched})^3$$

Note: The METEOR score can be reduced by the maximum of 50% if there are no bigram or longer matches ($Penalty = 0.5$)

Limitations

- **Formula for computing the score is not optimal**

  Because the forumla used for computing Penalty and METEOR score were manually crafted based on empirical tests, improvement on the parameters can still be made after more training on data

- **Limited modules are used in alignment**

  So far, the vanilla METEOR only supports three modules (exact, porter stem, synonyms) in the process of creating alignment. More modules that can measure semantic relatedness can be incorporated to produce better alignment result.

- **Ineffective use of multiple reference translations**

  METEOR comapres the machine translation with each reference translation separately and select the reference with the best match. The need for multiple references may diminish as better matching approaches are developed.

- **Identitcal weights of matches produced by different modules**

  METEOR gives the same weights to matches produced by different modules. Since different modules are

matching words by different semantic criterion, a better way might be applying different weights to the mapping produced by different mapping modules.

Reference

1. Papineni et al. BLEU: a Method for Automatic Evaluation of Machine Translation, ACL, 2002.
2. Banerjee et al. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments , ACL workshops, 2005.